COPY 2

OCR OPPORTUNITIES IN THE NATIONAL LIBRARY OF MEDICINE

Submitted to the Lister Hill National Center for Biomedical Communication

June 30, 1969

U. S. Department of Commerce

National Bureau of Standards

Center for Computer Sciences and Technology

OCR OPPORTUNITIES IN THE NATIONAL LIBRARY OF MEDICINE

Submitted to the Lister Hill National Center for Biomedical Communication

June 30, 1969

U. S. Department of Commerce

National Bureau of Standards

Center for Computer Sciences and Technology

Archi
Z
673.3
U58
1969
c. 2

OCR Opportunities in the National Library of Medicine

1. Introduction

OCR (optical character recognition) techniques offer significant promise of improved input in a variety of information processing applications, specifically including bibliographic announcement and control operations such as are found at the National Library of Medicine. For example, with respect to the MEDLARS (Medical Literature Analysis and Retrieval System) program at NLM, it has been said that: "While considerable improvement may be expected in basic keyboarding processes it appears from our perspective that the greatest potential for breaking the input 'bind' lies in optical scanning." (Lannon, 1967, p. 53).

The state-of-the-art in optical character recognition, both practical and experimental, is indeed promising, but many challenges still remain. Current success in terms of practical applications is largely limited to those cases where there is a high degree of control over character input quality, where the character sets to be recognized are limited (and often consist of specially designed character fonts), and where the alternative of key-stroking the input material is excessively costly in terms of available manpower and time.

In particular, application of OCR techniques for library and bibliographic processes presents special difficulties and specialized requirements. For example, "the cost of converting printed data already in libraries is still prohibitive. Practical conversion must await more economical character-reading machines and similar devices for encoding drawings." (Herbert, 1966, p. 32).

For another example, we note the following from the request for proposal for MEDLARS II ("Functional System Specifications for the National Library of Medicine"):

"Automatic printout of a book-form dictionary of definitions and scope notes, including chemicals, drugs, and synonyms, essentially the equivalent of publishing the dictionary card file. The main problem would be the conversion of the existing dictionary file." (p. 4-39)

Accordingly, and in the light of the NLM objective to seek "creative new solutions to library requirements", a study of OCR opportunities as they now exist in NLM or as they might be developed has been undertaken by personnel of the National Bureau of Standards at the request of the Associate Director for Research and Development for the Library, who is also the Director of the Lister Hill National Center for Biomedical Communications.

The obvious first question is: what present or proposed NLM tasks might benefit from available or developmental OCR techniques? The second question is very like, yet subtly different, namely; what available or potentially available OCR equipment could be of benefit to present or proposed NLM operations? The NBS study team has addressed itself to these questions in terms of requirements analysis, resources analysis, and cost-benefit considerations. The results to date will be discussed below, following a summary of our findings and recommendations. In addition, some further research and development requirements, involving advances in the state of the art of character and pattern recognition which may be of significance to <u>future</u> NLM applications, are briefly discussed.

2. Summary of Findings and Recommendations

The preliminary findings of the NBS study team were somewhat pessimistic as to the adoption of <u>available</u> OCR techniques based upon an estimated workload of approximately 100,000,000 characters per year, or 8,500,000 characters per month. At this level, the cost-benefit ratio for the introduction of a multifont page reader would appear marginal.

However, additional workload areas are involved in the MEDLARS II proposals including the Augmented MeSH Data Base, the Item Record Data Base, and in particular the conversion of up to 80,000 abstracts per year (averaging 1,000 characters each) to machine-readable form.

Furthermore, additional sources of supply of OCR equipment indicate the probable availability of multifont page reading capabilities at significantly lower cost. In addition, some OCR service bureau organizations are currently offering per-thousand-character rates significantly below the present rates (both in-house and on contract) of approximately $1.00 per 256 characters, or $4.00 per thousand.

A number of alternatives have been considered. These are discussed below in terms of relative advantages, disadvantages, and actions required in order to implement each alternative if adopted.

The alternatives are:

(1) Continuance of present input procedures and specifically the use of Flexowriters for Index Medicus and Current Catalog inputs.

Advantages. This alternative capitalizes upon present efficiencies, equipment and facilities.* Present costs and production rates are very reasonable, especially in view of the complexity of the character sets involved. The current

---

* An extended character set is available by modification of keys on the currently used machines.

method involves the typing of the journal identifier only once "and then the individual article is typed in using both the journal and the data forms." ("Functional System Specifications", p. 4-85).

Disadvantages. This alternative perpetuates current time-lags, such as those involved in copy correction, and would tend to prevent significant expansion of indexing or cataloging coverage.

Actions indicated. Added staff and/or contractual services would be required to handle backlogs and new items, such as the abstracts. (In the latter case, for example, up to 15 additional typists may be required).

(2) Continuance of present procedures for current workloads, but isolatable new tasks, specifically the preparation of the abstracts, to be processed by an OCR service bureau.

Advantages. This second alternative has the same advantages of (1) above but in addition it would presumably eliminate the need for additional NLM staff and it would provide an introduction to and a growing familiarity with OCR techniques for Library personnel.

Disadvantages. The disadvantages are the same as in alternative (1).

Actions indicated. It is suggested that an RFP be prepared and distributed to potential suppliers of OCR facilities on a service bureau basis. These would include, for example, Computer Optics and Scanning Corp., Wash., D.C.; Control Data Corp.; Farrington; Input Services, Inc., Dayton, Ohio; Source Data Automation Corp., Marlow Heights, Md., etc.*

(3) A third alternative is to proceed with the introduction of OCR techniques on the basis of a service bureau contract.

Advantages. The advantages of this alternative are:

- Gains in turnaround time, capacity for expanded workloads, and probably lower input costs with minimum disruption of current procedures and practice by both professional and clerical personnel.
- Presumably, significantly less cost for the addition of the abstract workload, since, on another problem (with far less "complexity" of input character set, however), bid prices per

---

* It is also possible that OCR services may be available on a reimbursable basis within Government at a later date.

thousand characters (i.e., precisely the expected average length of the abstracts) ranged from about $0.75 to slightly over $2.00. This is to be compared (with due regard to the differences in complexity) to the present rates of $1.00 , or more, both in-house and on contract, per maximum of 256 characters per citation.

- The service bureau approach would enable a relatively easy change-over to owned or leased equipment with more sophisticated capabilities at a later date.
- A service-bureau OCR commitment, especially for the added textual input for MEDLARS II of 1,000-2,000 character abstracts, has the following specific advantages:
    - (a) Minimal cost
    - (b) No capital investment, maintenance, or depreciation charges
    - (c) Throughput, quality control, and protection features required as necessary conditions of contract fulfillment
    - (d) NLM experience, and growing expertise, with this type of input.

Disadvantages. The material to be processed must be transported to and from another site, perhaps in a different geographic area. The character complexity of the material may be beyond the experience of the service bureau typists, resulting in a high error rate in the initial typing. On-line correction facilities would not be immediately available to NLM personnel. Backup facilities may not be adequate to assure continued production to meet publication deadlines. Character sets or fonts available with service bureau equipment may be inadequate for NLM purposes. It is probable that no advantage can be taken of NLM direct typing possibilities. Changes in pricing or scheduling policies might occur with inadequate advance notice.

Actions indicated. This alternative would require an NLM task force to carry out detailed and exhaustive analyses of specific requirements for each of the workload areas to be considered for immediate OCR processing as well as the preparation, distribution, and evaluation of responses to an appropriate RFP. Special attention should be paid to the following system design considerations:

- Requirements for the re-design of data input forms and formats.

- Possibilities for decentralization of input item preparation.

(4) Use of leased or purchased OCR equipment with programmable multifont capabilities and a minimum character set of 128 distinguishable characters for all major input processing operations.

Advantages. This fourth alternative offers many of the advantages of alternative (3) but with the added features of on-site availability, possibilities for on-line interaction as desired, and opportunities for extra-shift utilization. If, as is likely, an owned or leased OCR installation of the type recommended is not fully occupied with production operations, then

- The programmable features of format control should enable effective experimentation with:
  - The NLM development of appropriate edit/display routines
  - The extension of available character sets to include other character-types that are desired.
  - The desired provisions for hand-printed entries in given formats may be tested out.

. Additional information, from abstracts of items of chronological date earlier than that now contemplated in the MEDLARS II specifications, may be entered into the system.

In view of the programmable features, limited recognition of special identifiers, such as personally-hand-printed inputs, may provide important access-authentication checks.

Other advantages are that turnaround time --- from original input through initial processing to error indication, error correction, and re-entry of corrected data --- should be significantly reduced and that present problems of additional coverage --- in terms of lack of human resources and processing time --- could be alleviated to an important extent. Immediate advantage could be taken of existing direct typing, e.g., by indexers.

Disadvantages. The adoption of the full multifont OCR alternative might be prohibitively expensive in terms of capital investment or rentals, maintenance, and depreciation with respect to the benefits to be realized. Retraining and suitable motivation must be provided to both professional and clerical personnel in order for them to adjust to necessary changes in practices and procedures.

Actions indicated. An even more intensive requirements analysis effort than that required for alternative (3) is indicated. Personnel re-orientation and re-training must be planned and implemented. In addition to the system design considerations for (3), above, we note the possibilities for automatic proofreading of GRACE outputs, the possible requirements for new notational techniques, and requirements for quality control, including, for example, provisions for measurements of print quality.

In terms of economic advantages, the possibilities of joint financing of an OCR system might be explored with other constituents of HEW --- for example, the Clearinghouse for Mental Health Information, which has somewhat similar bibliographic control and processing problems. It should be recognized, from the outset, that a multifont machine capability installation, whether leased or purchased, may be under-utilized in terms of production operation requirements. Alternately, the possibilities of offering service bureau facilities, especially for off-hour use, might be considered.

(5) Use of leased or purchased OCR equipment of modular design, with initial capabilities for single-font reading of up to 128 distinguishable characters, and with additional font capabilities (including hand-printing) to be exploited at a later date.

Advantages. This alternative has many of the same advantages as alternative (4) above. Initial investment costs will be less and actual benefits can be checked out before major cost increments are committed. Modular design permits a gradualistic approach both in terms of application areas selected for implementation and in terms of personnel re-training and of forms re-design. On the other hand, additional fonts can be added to the system to meet further requirements, up to and including the direct reading of some journal pages.

The completion of one-time file conversion operations (such as the entire serial record for cataloged items from 1960 onward) would, of course, progressively free the equipment for expanding workloads --- e.g., from 12,900 titles cataloged in 1966 to 28,000 anticipated in 1972, or from 109,300 serial issues received in 1966 to the estimated 266,400 for 1972.

Disadvantages. The disadvantages of installation and conversion costs and of resistance to change should be considerably less than for alternative (4).

Actions indicated. The actions required to implement this alternative include those given in (4) above.*

(6) Conversion of present Flexowriter or keypunch input operations to the use of either stenotyper or of direct keyboard-to-magnetic-tape equipment.

Advantages. There has been some evidence that the use of either stenotype or magnetic tape typewriter equipment may show both cost reduction and productivity gains by comparison with other keyboard methods of input. For example, "In addition to providing instantly verified magnetic tapes, this ... [tape typewriter] system provides editing and retyping aids which may improve secretarial typing throughput up to a factor of 1.9." (Moore, 1967, p. 31); "If a stenowriter can be used as the input device, the production rate may be 4 times greater than that of typing." (Moore, 1967, p. 77). The use of "magnetic tape typewriters for conversion of data to machine readable form" is specifically recommended in the "Functional Systems Specifications."

---

* In this case it is likely that there will be a single source of supply, the Scan-Data Corporation (see Section 4 below).

Disadvantages. The disadvantages of this alternative are similar to those of alternative (4), but without the advantages of possibilities for present and future direct reading. Total costs per word have been estimated to be about the same for the magnetic tape typewriter and for re-typing for OCR input where 50 conversion personnel are required (Moore, 1967, p. 91), but with the less expensive multifont techniques now available, OCR costs per word should be less. The error rate for direct typing to magnetic tape is estimated to be 2.0 percent as against 0.9 percent for both OCR typing and flexotyping. (Moore, 1967). It is noted further that: "Magnetic tape encoders ... offer an alternative to keypunching, but the difficulty of inserting material at random restricts their application." (Van Dam and Michener, 1967, p. 189).

Actions indicated. Personnel re-training.

On the basis of the above findings, the NBS study team submits the following recommendations:

Recommendation 1. The National Library of Medicine should proceed with the necessary further requirements analysis and systems design pursuant to alternative (5) above provided that OCR equipment purchase can be limited to a cost of less than $400,000 and/or the equivalent in lease or rental arrangements.

Recommendation 2. An appropriate request-for-proposal should be prepared and submitted to known suppliers of multifont page reading OCR equipment, notably: Philco, Farrington, Control Data Corp., Op Scan, IBM, Recognition Equipment Inc., Mergenthaler-Linotype, Information International Inc., Compuscan, and Scan-Data Corp. The RFP should include in the mandatory requirements at least the following:

- Throughput costs, per thousand characters, not to exceed present costs.
- Programmable control for variable input* formats and for other purposes, compatibility with ASCII code and GRACE character set requirements, precedence or error detection and display, error correction inserts, and the like.

---

* For example, "With a stored-program controller, the system can determine three very important things during a single reading pass: (1) Whether or not there has been a mistake in data preparation, (2) Whether or not there has been an omission in data preparation, and (3) Whether or not the machine has read the data correctly. Also, the system can edit, accumulate, balance, verify check-digits, check parity, and condense data to provide easier access and reduced storage costs. Exception documents can be marked and sorted during the single reading pass, and details can be printed on a peripheral printer so that corrections can be made easily." (Philipson, 1966, p. 128).

- Capability for recognizing at least 128 character types, whether in single font or multifont (including handprinted versions).
- If a single font implementation meeting the other requirements, is initially proposed, the capability, by modular extension, of meeting multifont and handprinted requirements at a later date.
- Stand-by or back-up facilities, preferably on a service bureau basis.

Recommendation 3. In the event that responses to the RFP do not meet the above requirements (it is known that at least one potential supplier, the Scan-Data Corporation, can theoretically do so within or below the suggested price maximum), it is recommended that the service bureau approach for all or part of the present and proposed input processing as in alternatives (2) or (3) should be adopted.

These recommendations are submitted with the following caveats:

(1) It should not be assumed that the OCR installation, as presently available, would be capable of handling anything other than the high-volume typed or key-punched inputs, i.e., Index Medicus, Current Catalog, abstracts, and so forth (presumably, the handwritten entries would require re-typing for OCR, at present).

(2) The character set available will be minimal, but in accordance with MEDLARS II specifications in a single (or several closely related) font(s) upon installation.

(3) The outputs of either the OCR equipment, or the subsequent processor, or both, shall be ASCII-compatible or ASCII-convertible.

(4) The recommended equipment cannot be applied at this time to the solution of the problems of reading from microforms with a wide variety of fonts, type styles, and formats, or of recognizing complex graphic symbols such as chemical structure diagrams.

The NBS study team therefore also suggests:

Recommendation 4. The National Library of Medicine should support research and development efforts in such areas as the direct reading from microfilmed pages of representative journals (including automatic extraction of portions of text labelled "Abstract" or "Summary") and the automatic recognition of chemical symbols and diagrams.

3. Requirements Analysis

The first step in the OCR study was the preparation of a detailed plan of attack, stressing a systems engineering approach, as shown in Attachment 1. A first-cut estimation of probable OCR workload, however, indicated that for a probably marginal application (at the then estimated costs of equipment sufficiently powerful and versatile for NLM purposes) efforts requiring considerable time and NLM manpower should not be pursued at that time.

The situation has changed with the advent of multifont equipment with flexible character sets at significantly less cost. Hence, it would appear that a break-even point can be achieved for the following workload:

1. Catalog records, monographs --- 500 each two weeks, or 13,000 entries per year with 256 characters per entry and a correction factor of 0.58 --- 8×10^6 characters/year.
2. Indexing of periodicals --- 2,300 periodicals with 200,000 articles per year, 256 characters per entry, and 0.20 correction factor --- 64×10^6 characters/year.
3. Additional indexing of 600 periodicals per year --- $\frac{600}{2300} \times$ 64,000,000 characters per year = 16.7×10^6 characters/year.

In addition, the following workloads are directly anticipated in accordance with the MEDLARS II proposals:

1. Medical literature abstracts --- 80,000 per year with average length of 1,000 characters each and an estimated correction factor of 0.20 --- 96×10^6 characters/year.
2. Item record data (title, catalog, processing, holdings, routing, and usage data for material in the collections or under procurement) backlog from 1960 --- 150,000 items, variable length, 500 characters minimum assumed --- 90×10^6 characters one time, yearly load not estimated.
3. Augmented MeSH vocabulary --- a minimum conversion requirement of 9,000 scope notes at 425 characters each; 9,000 history notes at 425 characters, and 123,000 indexing instructions at 150 characters --- a one-time load of at least 26×10^6 characters, yearly increments not estimated.

Further, there are other potential workloads such as interlibrary loan requests (150,000 per year at 60 characters per record, or 9×10^6 characters per year), on-site reader requests (100,000 per year at 60 characters each, or 6×10^6 characters per year), and a cataloging backlog estimated at 24×10^6 characters.

Thus there is a potential initial workload (of one to two years duration) of not less than 240,000,000 characters per year, and 20,000,000 characters per month. This is well within the estimates for "re-typing for OCR" break-even thresholds (as discussed in Section 5 of this report). It should be stressed, however, that this conclusion with respect to requirements analysis is based upon the assumption that a number of MEDLARS II proposals will in fact be adopted.

On the other hand, in practice, advantage can and should be taken of present direct typing --- whether by indexers, catalogers, or other personnel preparing orders, invoices, dictionary cards, category lists, new medical subject headings, and the like. Moreover, the Scan-Data equipment which would meet the suggested RFP requirements will have hand-print recognition capability either on initial installation or for subsequent implementation.

Some pertinent factors that were brought out in discussions with NLM personnel are as follows:

- There is some dissatisfaction with present methods of input. In particular, there are scheduling difficulties with input proof corrections and turn-around times in general are too slow.
- Desirable increases in coverage, both of monographs and journal titles, are limited by lack of both indexing and input resources.

- Typing resources available to OCES, in-house and on contract, amount to the equivalent of 25 typists, with three more needed for the current workload, and space is at a premium.
- Some of the material prepared by indexers and catalogers is re-typed by input flexotypists. This includes, for example, MeSH headings, transliterations of titles, translations of titles, and some corrections.
- About 50 percent of the indexing work is reviewed with additions and deletions indicated by pen or pencil. This handwritten information is not likely to be machine-interpretable.
- It apparently would not be too difficult to change individual typewriters in the technical divisions, such as the 15 used by catalogers in the Technical Services Division.
- There are problems with punched paper tape, but on the other hand the flexotypist is able to carry journal and issue code along for each article in each issue by use of a special stroke. (This could also be accomplished with an OCR stored program).

- A special problem may arise if OCR techniques are adopted in the case of information that is recorded by rubber stamp.
- It may be desirable to conduct experiments with microfilming for OCR reading, looking toward the ultimate conversion of 900,000,000 cards (See Recommendation 4).
- The Reference Services Division would give high priority to mechanization of loan transactions and reader service records for management information purposes.
- Since approximately 50 percent of the literature processed is in languages other than English, the character set is complex.

The last of the above factors points to a special consideration: namely, that at a current rate of 4,000 characters per hour or approximately 13 words per minute (Lannon, 1967) and high complexity (see Moore, 1967, p. 50-51), there is likely to be less productivity gain from the introduction of OCR techniques than might otherwise be the case. Nevertheless, major gains can be expected with respect to:

- Improved turnaround times, including error processing after proofreading or computer rejects.
- Capacity for increases in workloads, for development of additional applications within NLM, and/or for sharing of facilities with other organizations on a scheduled basis.

4. Resources Analysis

In the area of analysis of available and potential resources, the NBS team has reviewed the current state of the art of optical character recognition, with emphasis upon multifont page reading capabilities, microform-input, and reading of handprinted materials.

There are a number of potential suppliers of OCR equipment of varying levels of sophistication, capability, and performance, as shown in a chart prepared by Standard Register, a copy of which is provided as Attachment 3. Relatively few of these, however, have the character set capacities likely to be required in any NLM application; even if limited to a single font.

Presently available approaches involving machine reading techniques that could be applied to the Library's input tasks are: (a) typing in an OCR-acceptable font and character set, proofing as required, and machine reading to magnetic tape; (b) handprinting within designated constraints (such as the use of boxes or dots printed in "drop-out" ink, and the like), followed by microfilming and machine reading to magnetic tape, and (c) using a combination of typed and handprinted inputs to a reader.

The equipment of 10 manufacturers known to have actual or potential capabilities for reading hand-printed material has been investigated. Of these, four (Philco, Farrington, Op Scan, and IBM) do not offer equipment with sufficient sophistication for the NLM problems (e.g., limited

character sets in general and with particular respect to hand-printing). The CDC 915 is similarly limited, but a much more powerful CDC machine will soon be available (i.e., July, 1969).

On-site inspections have been made for the following multifont equipments: CDC 915 in use at McDonnel Douglas Corporation; Recognition Equipment, Inc.; Mergenthaler-Linotype, Inc., Information International, Inc., and the Scan-Data Corporation. In particular, the field trip report for the study team's visit to the Scan-Data Corporation is of special interest and is given as Attachment 2.

Recognition Equipment, Inc., (REI) has a type (c) approach (e.g., for an application in the Library of Congress) but will have only a one or two line per document capability for the near future. Mergenthaler-Linotype and Information International, Inc., both exhibited equipment that warrants serious consideration for <u>future</u> microform reading. Compuscan is a new organization capitalizing on prior experience with the Mergenthaler-Linotype approach.

There is no evidence of any immediate gain to be achieved (e.g., the next 12 months) by the use of microforms for <u>current</u> inputs. Furthermore, highly variable fonts, formats, and graphic interpolations involved in microfilmed material from the permanent collections (as shown in NBS Report 9446, "Report of a Study of Requirements and Specifications for Serial and Monograph Microrecording for the National Library of Medicine", a copy of which is attached to the original of this report)

indicate that considerable further effort both by the potential supplier(s) and by the Library would be required to actualize this possibility.

A recapitulation of characteristics of multifont reading systems potentially suitable for NLM applications is shown on the next page.

Some further details with respect to the Scan-Data equipment are as follows:

- The machine typically has full capability (multifont, 800 character/second reading speed, etc.) when built, but in effect is "disabled" back to minimum configuration to meet customer requirements.
- Additional character sets (100-150 characters possible) can be easily added at the field site.
- Five character sets are now available, i.e., OCR A, OCR B, Elite 10-pitch, Elite 12-pitch, and 1403 upper case.
- Character sets planned include other typewriter (10- and 12-pitch) fonts and typeset fonts such as Univers, Roman, and Gothic (the latter currently demonstrable) as well as hand-printed alphanumerics.
- Prices for a configuration to meet the requirements of our recommendation were quoted June 26, 1969 as follows:

Firm	Control Data Corp. Rockville, Md.	Compuscan Leonia, N. J.	Mergenthaler Linotype Plainview, N. Y.	Information International, Inc., Boston, Mass.	Scan-Data Norristown, Pa.
Fonts available in production machine	7 - can be extended to 20-50 later	many - 6 at present intermixed	8 - fonts not designated as yet	family of fonts, including commonly used typewriter	1-5 at present
Reading rates	14,000 char/sec.	microfilm read, equiv. 2,000 char/ sec.	300 char/sec. production	2,000 char/ sec. design goal	400 char/ sec. 800 char/ sec.
Price	$1.5 x 10^6	$0.9 x 10^6 estimated	$0.5 x 10^6 estimated	$1.2/15 x 10^6	$0.25 x 10^6
Delivery from date of order	18 mos.	10/12 mos.	12/18 mos.	1st machine early CY 70 delivery	6/8 mos.
Service Bureau	Yes	Yes 9/1/69	Not known at present	Yes, CY 70	Yes, West Coast
Hand-print	If desired	If desired	Not known	Numeric only at present	If desired

Basic Machine -	$140,000
Control Computer and Tape Deck -	$ 20,000
	$ 48,000 - 7 channel
-	$ 54,000 - 9 channel
On-Line Display (for error and reject correction) -	$ 12,000 (optional)
1 Character Set -	$ 30,000
	$256,000

- Each added character set - $ 30,000
- Delivery is usually 6 to 8 months after date of receipt of order, depending upon prior order scheduling. Thus, currently, there are opportunities available for December, 1969; January, 1970, and after June, 1970.

5. Cost-Benefit Considerations

The most significant factor with respect to our recommendations is that of cost as compared to workload and to anticipated benefits. In terms of our initial evaluation of the prospects for introducing OCR techniques into NLM operations, considerable attention was paid to suggested break-

even considerations such as those proposed by W. Moore of the Rome Air Development Center. Specifically, "Because of the high cost per word, it is not feasible to select either optical character recognition or entry/display complexes as a conversion method for a file being converted at a rate lower than approximately eight million characters per month." Moore reports further, however, that "an independent study indicates that this cutoff point may be as high as 16 million characters per month."

At that time, an estimated annual workload of 103.7 million characters per year (not including the preparation of 80,000 abstracts in machine-usable form) indicated that direct typing or re-typing for OCR would be marginal in terms of cost-benefit considerations. However, the Moore data was based upon the assumption of a capital investment cost of $530,000 (1967 estimate for procurement of a Philco page reader) for OCR equipment and, by coincidence (accidental or otherwise), the $256,000 price quoted by Scan-Data (in June, 1969) is not quite half this estimated cost. Conservatively applying this cost reduction factor, we find the following assuming only a one-third reduction of "input terminal costs":

	Cost in cents per word	Words per month (millions)
(For 25 input preparation or file conversion personnel):		
Flex.	0.3850	2.99
Mag. Tape.	0.4227	3.17
OCR	0.6507	4.40
Revised OCR	0.5117	4.40
(For 50 input preparation or file conversion personnel):		
Flex.	0.3912	5.85
Mag. Tape.	0.4335	6.21
OCR	0.4466	8.60
Revised OCR	0.3771	8.60

Assuming a minimum NLM input preparation and file conversion staff of 40, (25 to 28 for current OCES operations, but 11 catalogers and 35 indexers do some typing) we find that the cost factor for the less expensive multifont OCR approach is reasonable.

Among the many added benefits are decreased turn-around times and increased productivity such as to enable the addition of the abstract preparation and other application tasks. We may also note the following:

"One may readily ask: 'If the input data must be rekeyed, what is the advantage of ... optical scanning?' The answer lies in the fact that many typewriters equipped with normal type font can be readily changed to ... [a pre-selected] optical font by mere selection. The ordinary typewriter can then become a substitute key punch device. The potential advantages of a page scanner are:

"1. The input keying of the library surrogate can become decentralized. The elements of the surrogate can be typed on a document 'traveler' and added by one station after another, the final station performing the final editing on the surrogate.

"2. Hidden codes become non-existent. What the proofreader reads on the document is what will be read by the computer.

"3. The difficulties in creating a batch for the computer to process are removed. Selected pieces of paper can themselves be made into a batch, and no coordination of visual record and paper tape rolls is required." (Wishner, 1965,

6. Some R & D Considerations

An alternative not previously considered in this report is that of on-line input and recognition via personal terminals for general computer interaction where the main processor can be "taught" to recognize a variety of characters and symbols, including those that are unique to a particular individual. It is our feeling that at this time such an approach would require considerable R & D effort with respect to NLM users and their requirements. Further, it would appear that implementation of such an approach must await the development of the final system capabilities for MEDLARS II.

As has been noted previously, microform recognition techniques under development may have a significant future potential for NLM operations. Thus, hand-printed items of various sizes and formats could be microfilmed for scanning. Redesign of forms could be held to a reasonable minimum. These experimental capabilities might also be applied to the reading of currently existing microfilm. Of particular interest would be the solution of paper handling problems in the scanner by the use of microfilming, although some attention must be given to this step at the microfilm camera.

Two of the suppliers investigated, III and Compuscan, would probably be receptive to an R & D contract or subcontract proposal to investigate microfilm input potentialities. Compuscan in particular offers service bureau facilities. In addition, either or both organizations might be amenable to the undertaking of R & D tasks in connection with the processing of hand-input or preprinted chemical structure information including diagrams. In the latter case, however, it must be emphasized that a considerable amount of time must be devoted to the task by trained chemists and other specialists thoroughly familiar with NLM requirements.

7. Conclusion

It is concluded that OCR equipment of the type represented by Scan-Data could be effectively used in NLM operations for a period of at least three to five years, which would enable amortization if purchased outright. It is likely that there would be some substantial continuing workload (including inputs from international collaborators) even after on-line indexing and editing stations might come into use or if developments in microform recognition processing and graphic recognition should dictate a shift to such more advanced equipment.

NBS personnel will be pleased to render any further assistance to NLM as may be requested, whether for the requisite further requirements analysis, systems engineering with particular reference to a number of new interfaces, forms re-design, procurement and installation, initial operation, and/or for the suggested experimentation.

REFERENCES

Forbes, E.J. and T.C. Bagg, Report of A Study of Requirements and Specifications for Serial and Monograph Microrecording for the National Library of Medicine, unpublished report, National Bureau of Standards, Aug. 1966, 100 p.

Herbert, E., Information Transfer, Int. Sci. & Tech. 51, 26-37 (1966).

Lannon, E.R., Optical Character Recognition in the U.S. Government, in Advances in Computer Typesetting, Proc. Int. Computer Typesetting Conf., Sussex, England, July 14-18, 1966, Ed. W.P. Jaspert, pp. 48-53 (The Institute of Printing, London, 1967).

Moore, W.B., The Input Problem, preprint of paper presented to the Washington, D.C., Chapter, The Institute of Management Sciences, Washington, D.C., Oct. 11, 1967, 107 p.

National Library of Medicine, Functional System Specifications for the National Library of Medicine, 1 v. (National Library of Medicine, Bethesda, Md., July 1, 1967).

Philipson, H.L. Jr., Optical Character Recognition: The Input Answer, in Data Processing, Vol. X, Proc. 1966 Int. Data Processing Conf., Chicago, Ill., June 21-24, 1966, pp. 119-130 (Data Processing Management Assoc., 1966).

Van Dam, A. and J.C. Michener, Hardware Developments and Product Announcements, in Annual Review of Information Science and Technology, Vol. 2, Ed. C.A. Cuadra, pp. 187-222 (Interscience Pub., New York, 1967).

Wishner, R., The Role of Paper Tape and Optical Scanning Computer Input in Textual Data Processing, Proc. 1965 Congress F.I.D. 31st Meeting and Congress, Vol. II, Washington, D.C., Oct. 7-16, 1965, pp. 235-240 (Spartan Books, Washington, D.C., 1966).

ATTACHMENT 1. Plan of Attack, OCR Program Planning and Development for NLM

Required Actions	Scope and/or Nature of Coverage	Output Products	Remarks
1. Development of detailed plan of attack and preliminary scheduling/ costing	1.1. MEDLARS inputs; 1.2. Toxicology; 1.3. CA; 1.4. Structure diagrams; 1.5. Microfiche; and 1.6. Other possible future desiderata.	Input to subsequent and concurrent actions.	
2. Planning of necessary fact-finding investigations.	2.1, System requirements; 2.2, availability considerations; 2.3, potentiality considerations; 2.4, feasibility considerations.	Determination of factors to be considered; development of checklists or interview structures; scheduling; input to subsequent actions.	2.3 includes, e.g., possible benefits from promising R & D efforts, potential added requirements for future expansions, contributions to provacy/confidentiality requirements, etc.
2.1 Development of System Requirements Factors	2.1.1, processing requirements; 2.1.2, performance and reliability requirements; 2.1.3, quality control mechanisms; 2.1.4, future expansion and flexibility requirements.	"	
2.1.1 Processing Requirements	2.1.1.1, Inputs; 2.1.1.2, processing operations; 2.1.1.3, outputs; 2.1.1.4, reject handling.	"	
2.1.1.1 Inputs (example)	Nature of input items (i.e., printed catalog cards, bibliographic references, typed or handprinted abstracts; well-structured drawings or diagrams, pictorial and photographic data, etc.); carriers (paper, microforms) by type;		

Required Actions	Scope and/or Nature of Coverage	Output Products	Remarks
	system design characteristics for each type; physical characteristics by type (dimensions, paper quality, reduction ratio, etc.); feed requirements by type; throughput speed requirements by type; volume and percent of total volume by type; formats/type and whether or not controlled; fonts (type, number, frequency of occurrence); character sets (number, type, size per type, frequency of usage by type, etc.); typical expected prior error or noise by item type, etc.).		
2.1.2 Performance Requirements	2.1.2.1, throughput; 2.1.2.2, uptime, maintenance and standby; 2.1.2.3, error tolerance; 2.1.2.4, reject tolerance; 2.1.2.5, system component redundancy.	"	
2.1.3 Quality Control Considerations	2.1.3.1, preparation, input, processing, output, and feedback controls available; 2.1.3.2, feasible additional controls.	"	
2.1.3.1 (example)	Quality control measures available for: quality of paper or other carrier media; recording or transcription procedural requirements (e.g., no strikeovers permitted); verification of recordings or transcriptions; imprinting (e.g., uniformity of ink density); format; format variety; registration (e.g., tolerances for item, line and/or character skew; use of fiduciary marks); limitation		

Required Actions	Scope and/or Nature of Coverage	Output Products	Remarks
	to specialized font(s); use of restricted character sets; context dependent restrictions (e.g., alpha or only numeric); item variety mix; carrier variety mix (e.g., whether paper records of varying physical dimensions and quality can be presorted or must be processed intermixed), font mix (including text-graphic interpositions), etc.		
2.1.4 Expansion and Flexibility	2.1.4.1, added volume; 2.1.4.2, additional kinds of input required; 2.1.4.3, additional processing required; 2.1.4.4, additional kinds of output; 2.1.4.5, added outlets.	"	
2.2 Development of Availability Consideration Factors	2.2.1, minimum specifications, equipment physical characteristics; 2.2.2, minimum specifications, equipment performance characteristics; 2.2.3, system compatibility/convertibility factors; 2.2.4, availability and costs for purchase, lease, or rental; 2.2.5, service bureau availability; 2.2.6, maintenance and repair considerations.	"	(E.G., 2.2.1 might include required resolution, per input item and carrier type.)
2.3 Development of Potentiality Consideration Factors	2.3.1, advanced techniques of error detection and correction; 2.3.2, advanced hardware developments; 2.3.3, advanced pattern recognition developments; 2.3.4, theoretical pattern recognition research.	"	Possible benefits from promising R & D efforts would be appraised in terms of estimated likeliness of success, time-scale of possible success, feasibility of production engineering, costs of R & D contributions

Required Actions	Scope and/or Nature of Coverage	Output Products	Remarks
2.4 Development of Feasibility Criteria	2.4.1, comparative benefits, costs, alternative methods; 2.4.2, questions of centralization-decentralization; 2.4.3, difficulties, time, costs of conversion; 2.4.4, problems of re-formatting; 2.4.5, human factors, etc.	"	
3. Identification of likely sources of information	3.1, key personnel, MEDLARS toxicology, etc.; 3.2, professional and trade literature; 3.3, potential suppliers; 3.4, present and potential users; 3.5, evaluators (e.g., Auerbach reviewers, RADC personnel).	Interview lists and schedule.	
4. Layout of forms for tabulations and analyses and for interview check-lists and/or questionnaires as indicated	2.1, 2.2, 2.3, 2.4 above.	Work sheets	
5. Conduct of fact-finding investigation	2.1, 2.2, 2.3, 2.4 above.	Data for analysis and evaluation.	D. Friedman and F. Wirdzek of NBS will check on microform reader possibilities, RCA and Mergenthaler, new few weeks.
6. Analysis and evaluation of findings	6.1, By area of possible application; 6.2, by type of input item within area.		

Required Actions	Scope and/or Nature of Coverage	Output Products	Remarks
7. Development of specific recommendations	7.1, Further system design requirements analysis, if necessary; 7.2, implementation, to the extend indicated, of available techniques; 7.3, independent or collateral support of R & D developments, if indicated.	Recommendations, supporting data, and evaluation report.	
7.1 System design requirements	7.1.1, What; 7.1.2, why; 7.1.3, who; 7.1.4, when; 7.1.5, how.		
7.2 Recommendations on available techniques	7.2.1, Specifications and RFB; 7.2.2, evaluation and selection; 7.2.3, purchase, lease, or rental and installation or service bureau contract; 7.2.4, pilot or dual operation; 7.2.5, transfer to full production status, if indicated.		
7.3 Recommendations for R & D support	7.3.1, What and why; 7.3.2, who; 7.3.3, how much; 7.3.4, how to support.		Probable emphasis on: (a) constrained hand-printed characters, (b) well-structured diagrams, (c) microfiche reading.

ATTACHMENT 2.

U.S. DEPARTMENT OF COMMERCE
National Bureau of Standards
Washington, D.C. 20234

Date: June 24, 1969

Reply to Attn of: 610.0

Subject: Trip Report - Scan-Data Corp.

To: File

An information gathering trip was made on May 27, 1969 to the Scan-Data Corp., 800 East Main Street, Norristown, Pa. 19401. The purpose was to review capabilities of page readers manufactured by this company and to assess their possible usefulness to Federal Government departments and agencies. Members of the visitation party included M. E. Stevens, Office of the Director, CCST; David C. Friedman, Div. 650 and Roy Worrol, Div. 640; and the writer.

Paper Handling

Demonstrations were made of the Scan-Data 200 Page Reading System. The machine includes the normal paper handling system components of Input Hopper, Paper Transport, and Output Hopper. The input and doubles feed control uses a pair of precision, metered, counter-rotating, plastic feed rolls. Paper transport is on a vacuum belt of unique material and construction. A skew measuring device is incorporated and misaligned pages are routed to a separate reject stocker(Three output stackers are used in place of the usual complement of two). No attempt is made to adjust for paper skew once the form has left the Input Hopper.

Paper feed is under program control using a precision stepping motor so that very small plus or minus feed increments (about 1/2 character height coarse feed or 5 mils fine feed) can be achieved. Stepping to accommodate a slightly skewed or wavy line is provided thru program control. Forms to be fed cannot be intermixed as to size or format. Different thicknesses in the same input stack can give feeding problems.

Scan System

Scanning is by a 10" cathode ray tube with 2 mil diameter scanning beam, generally at 1 tol magnification, using a P16 phosphor (near ultra violet) for early production models. Later models will have a phosphor which will bring the scan band in the yellow green region of the visible light spectrum. Four photo multiplier tubes are used to collect reflected light. Scanning is under software program control and is confined to steps on geometric, X-Y coordinate axes (no rotation of the scan field is used for correction of skew). Skew can be accommodated up to 1%, beyond this the form is rejected to the skew reject hopper. Scan rate is 400 characters/second

and will be increased to 800 characters/second in later models.

Recognition Logic

Signal processing is achieved in several logic steps with minor attention to noise clean up (suppression of stray dirt noise, filling of voids, or broken edges). Emphasis is on detection of "features" - approximately 300, consisting of 4 to 30 bits of information - found in sub-areas of character images (examples include the bar for capitol G, tail of the y, downward points of W, cross bar for lower case f, or the cross bar on t). Both presence and absence of required features are checked, and at present a "perfect" match is required in the recognition correlations. This requirement, however, can be relaxed to accommodate less perfect copy.

Character shapes may or may not be normalized. A clear but narrow vertical band is required between characters (overhanging characters are a problem) but abutting serifs can be suitably handled. There is a considerable variation in sizes of fonts that can be recognized.

Control Computer & Output

A PDP-8 or PDP-8I process control computer is used for edit and software control tasks. Program loading is via punched paper tape. System output is read to magnetic tape using whatever code system may be ordered by customers. Apparently there is no enthusiasm for using ASCII output codes unless specifically demanded.

The computer is used extensively for hardware control tasks such as selection of expected font at each field to improve the reading process, stepping for skewed lines, separation of characters, and presentation of errors for manual correction via an optional display and console unit. It also may be used to control the threshold level for the scanning process, and the normalization of scanned characters.

Demonstration

A variety of program documents were demonstrated in both formatted and unformatted page form layouts. The test forms were offset printed or robot typed with one-time ribbons and represented "perfect" copy. Several tests were made to determine effects of creases, wrinkles, strikeovers, dirt, confusion, etc. The system responded with acceptable performance.

Examples of OCR-B and Bell Gothic were read by a machine for R. R. Donnelly, demonstrated in the engineering test room. The OCR-B was read as numerics only or upper case only with Oh and Zero paired as a single character. No problem was found with Oh/D or Zero/D pairs.

Equipment Models

The SCAN-DATA-200 will also be available in an 800 character/second model. Multifont capabilities are provided in the SCAN-DATA-100 Model.

One machine is available at the Scan Data West Coast (Beverly Hills, Cal.) installation for service bureau work.

Max A. Butterfield

cc:
MEStevens
DCFriedman
RWWorral
JOHarrison, Jr.

MODEL & TYPE	CHARACTER READ	MARK READ	FONT STYLES READ	CHARACTER SET	SCANNING METHOD	USUAL IMPRESSING METHODS	APPLICATIONS	READING SPEED
ADDRESSOGRAPH 9600 OPTICAL CODE READER	No	No	A.M. Five Level Binary Code	Bar Code	Photocell	Imprinter	Credit Charging Petroleum Retail Hospitals	Up to 230 Cards Per Minute
CONTROL DATA CORP. 915 PAGE READER	Yes	Mark Sense Circles	915 Version of USASCSOCR	Alphanumeric Plus Symbols	Character Analysis by Photocell	Typewriter Pencil(Mark Read)	Updating of Files Subscriptions Addresses Status Changes	Up to 370 Characters Per Second
CONTROL DATA CORP. 935 DOCUMENT READER	Yes	Yes	915 Version of USASCSOCR, IBM 1428, 1428E, 407-1 Selfchek 7B & 12F	Alphanumeric Plus Symbols	Character Analysis by Photocell	Typewriter High Speed Printer Pencil (Mark Read)	Travel Tickets & Turn Around Documents	Up to 750 Characters Per Second
CUMMINS-CHICAGO ODPS 216	No	Yes	A.M. Five Level Binary Code, Binary One Code & Perforated Codes	Bar & Special Code	Photocell	High Speed Printer Imprinter or Cummins Perforators	Turn Around Docs. Invoices Payment Coupons Banking	Up to 500 Documents Per Minute
FARRINGTON 2030 PAGE READER	Yes	No	USASCSOCR Selfchek 12F & 12L	Alphanumeric Plus Symbols	Scanning Disc	Typewriter	Updating of Files Subscriptions Addresses Status Changes	Up to 400 Characters Per Second
FARRINGTON 3010 DOCUMENT READER	Yes	Yes	USASCSOCR Selfchek 12F, 12L & 7B IBM 1428	Alphanumeric Plus Symbols	Scanning Disc	High Speed Printer Pencil(Mark Read)	Turn Around Docs. Billing Sales Receipts Inventory	Up to 440 Documents Per Minute
FARRINGTON 3020/3022 CARD READER PUNCH	Yes	Mark Guide Circles	USASCSOCR Selfchek 12F, 12L & 7B IBM 1428 & 1428E	Numeric Plus E&P	Scanning Disc	Imprinter Typewriter High Speed Printer Pencil(Mark Read)	Credit Charging Petroleum Retail Hospitals	Up to 500 Cards Per Minute
FARRINGTON 3030 PAGE READER	Yes	Mark Guide Circles	USASCSOCR Selfchek 12F & 12L	Alphanumeric Plus Symbols	Scanning Disc	Typewriter	Updating of Files Subscriptions Addresses Status Changes	Up to 400 Characters Per Second
FARRINGTON 3040 TAPE READER	Yes	No	USASCSOCR Selfchek 12F, 12L, IBM 1428 & NCR NOF	Numeric Plus Alpha Control Symbols	Flying Spot	Cash Register Adding Machine etc.	Register Sales & Inventory	Up to 1000 Characters Per Second
G.E. DRD 200 BAR FONT READER	Yes	Yes	G.E. COC-5 Bar Font	Numeric	Photocell	High Speed Printer	Banking Payment Coupons Accounts Receivable	Up to 2400 Characters Per Second
IBM 1230, 1231 & 1232 PAGE READERS	No	Yes	Mark Reading Only	None	Photocell	Pencil(Mark Read Only) High Speed Printer	School Grading Inventory Sales & Status Reporting	1230 - 750/hr. 1231 -2000/hr. 1232 -1450/hr. Maximum
IBM 1282 CARD READER PUNCH	Yes	Yes	IBM 1428 & 1428E Selfchek 7B	Numeric	Scanning Disc	Imprinter Typewriter Pencil(Mark Read)	Credit Charging Petroleum Retail Hospitals	Up to 200 Cards Per Minute
IBM 1285 TAPE READER	Yes	No	IBM 1428 NCR NOF	Numeric	Flying Spot	Cash Register Adding Machine etc.	Register Sales & Inventory	Up to 540 Characters Per Second
IBM 1287 DOCUMENT READER	Yes	Yes	USASCSOCR IBM 1428 - 1428E SELFCHEK 7B NOF Handprinting	Alphanumeric (machine) + Symbols CSTXZ Numeric Hand-printing	Flying Spot	Imprinter High Speed Printer Typewriter Handprinting Cash Register	Sales Receipts Turn Around Docs. Inventory Billing	Depending on Form Design
IBM 1288 PAGE READER	Yes	Yes	USASCSOCR Handprinting	Alphanumeric (Machine) Numeric Hand-Printing +CSTXZ	Flying Spot	Typewriter High Speed Printer Handprinting	Sales & Inventory Reporting Updating Files	Depending on Forms Design
IBM 1418 DOCUMENT READER	Yes	Yes	IBM 407 & 407E-1	Numeric Plus Symbols	Scanning Disc	High Speed Printer Pencil (Mark Read)	Turn Around Docs. Billing Inventory	Up to 420 Documents Per Minute
IBM 1428 DOCUMENT READER	Yes	Yes	IBM 1428	Alphameric (Plus Symbols)	Scanning Disc	High Speed Printer Typewriter Pencil (Mark Read)	Updating Files Subscriptions Addresses	Up to 400 Documents Per Minute
MINN.-HONEYWELL ORTHOSCANNER 289-8	No	Yes (Bar Code)	H 1800 Hexadecimal Code	Bar Code	Photocell	High Speed Printer Pencil (Mark Read)	Utility Billing Insurance Payment Coupons	10 Char./sec. (Possible Variation to meet specifi application)
NCR 420-2 TAPE READER	Yes	No	NCR-NOF	Numeric Plus Symbols	Photocell	Cash Register Adding Machine etc.	Register Sales Inventory	Up to 3120 Lines Per Minute
OPSCAN 100 & 70 PAGE READERS	No	Yes	Mark Reading Only	None	Photocell	Pencil (Mark Read Only) High Speed Printer	School Grading Inventory Sales & Status Reporting	Up to 2500 Pages Per Hour
OPSCAN 288 DOCUMENT READER	Yes	No	USASCSOCR, E-13B IBM 1428, 407E Handprinting (Choice of One)	Numeric Plus CNSTXZ + and Hyphen	Photocell	High Speed Printer Typewriter Imprinter Handprinting	Sales Receipts Turn Around Docs. Inventory Billing	Up to 800 Characters Per Second Machine
PHILCO 6000 PAGE READER	Yes	Yes	Multifont	Alphanumeric Plus Symbols	Flying Spot	Typewriter Pencil (Mark Read)	Updating Files Invoicing Shipping	Up to 2000 Characters Per Second
RCA VIDEOSCAN DOCUMENT READER	Yes	Yes	RCA N-2	Numeric Plus Symbols	Vidicon Recognition	High Speed Printer Pencil (Mark Read)	Turn Around Docs. Billing Inventory	Up to 1500 Characters Per Second
REI ELECTRONIC RETINA DOCUMENT READER	Yes	Yes	Multifont Handprinting	Alphanumeric Plus Symbols	Photocell - Regina	Imprinter Typewriter High Speed Printer Handprinting	Turn Around Docs. Airline Tickets Petroleum Charges	Up to 2460 Characters Per Second
REI ELECTRONIC RETINA PAGE READER	Yes	Yes	Multifont	Alphanumeric Plus Symbols	Photocell - Retina	Typewriter High Speed Printer Pencil (Mark Read)	Updating Files Subscriptions Status Changes Airline	Up to 2460 Characters Per Second
REMINGTON-RAND CARD READER PUNCH	No	Yes	Mark Reading Only	None	Photocell	Pencil (Mark Read)	School Grading Inventory Status & Sales Reporting	Up to 9000 Cards Per Hour
SCAN DATA SERIES 300 PAGE READER	Yes	No	Multifont Handprinting	Alphanumeric Plus Symbols	Flying Spot	Typewriter High Speed Printer Handprinting	Insurance Claims Ordering Inventory Updating Files	Up to 400 Characters Per Second
MOTOROLA MDR-1000 DOCUMENT READER	No	Yes	Mark Reading and Hollerith Punching	None	Photocell	Typewriter (Mark Read Only) High Speed Printer Pencil	Insurance Claims Order Entry Billing Meter Reading	Depending on form length
HEWLETT-PACKARD 2760 & 2761 TAB CARD READER	No	Yes	Mark Reading and Hollerith Punching	None	Photocell	Typewriter (Mark Read Only) High Speed Printer Pencil	Inventory Order Entry Billing Meter Reading	Up to 105 Columns Per Second

DOCUMENT SIZES	MAXIMUM CHARACTERS PER LINE	PAPER WEIGHT RANGE	MACHINE FLEXIBILITY	OPERATING CONTROL	OUTPUT	SPECIAL FEATURES
Standard 51or 80 Column Tab Card	68	100# Tab Card Stock	Reads Selective Fields	Off Line	Punched Cards or Paper Tape	
4 x 2-1/2 to 12 x 14	110	15# to 100#	Reads Selective Fields under Computer Program Control	On Line with CDC 3000, 6000 and 8000 Series Computers	Data to Computer Punched Card Punched Paper Tape or Magnetic Tape	Reads Mark Sense Circles (Hand Filled)
3 x 2-1/4 to 5-1/2 x 8-1/2	80	20# to 125#	Reads Selective Fields	On Line CDC 1700	Data to Computer Punched Card Punched Paper Tape or Magnetic Tape	Batch Lister Control
4-1/4 x 2-1/4 to 8-3/4 x 4	82	24# to 100#	Reads Selective Fields	Off Line	Punched Paper Tape Magnetic Tape	
4-1/2 x 5-5/8 to 8-1/2 x 13-1/2	75	20# to 28#	Format Control by Plugboard Reads Selective Fields	Off Line	Punched Card Punched Paper Tape Magnetic Tape	Underscore Feature permits encoding of upper & lower case characters in output record.
2 x 2-1/4 to 6 x 8-1/2	64	24# to 125#	Format Control by Plugboard Reads Selective Fields	On or Off Line	Data to Computer Punched Cards Punched Paper Tape Magnetic Tape	Batch Header Mark Sense Head & List Printer Optional
Standard 51 or 80 Column Tab Card	65	100# Tab Card Stock	Format Control by Plugboard Limited Selectivity	Off Line	Punched Cards	Batch Header Serial & Sequential Numbering Reads Reverse Images
4-1/2 x 5-5/8 to 8-1/2 x 13-1/2	75	20# to 28#	Reads Selective Fields; Formating and Editing Facilities Provided	On Line with DMI620 Computer	Computer Punched Cards Punched Paper Tapes Magnetic Tape	Reads Mark Sense Accumulates Totals Formating & Editing
Standard Journal Tapes 1.31 to 3-1/4	32	Standard Journal Tapes	Format Control by Plugboard or External Computer Program	On or Off Line	Data to Computer Magnetic Tape	Journal Tape Header Entry Magnetic Tape Label Entry
2-1/2 x 5-1/2 to 3-3/4 x 9	50	20# to 100#	No Format Control Limited Field Selectivity	On or Off Line with any Computer	Data to Computers Punched Cards or Tapes Magnetic Tape	
8-1/2 x 11	1000 Total Response Positions Available	20# or 24# But Cal. .0045 to .0050	Reads Selective Fields	1230 - Off Line 1231 - On Line 1232 - Off Line	1230 - Score Printed on Form 1231 - Data to Computer 1230 - Punched Cards	
Standard 51 or 80 Column Tab Card	32	100# Tab Card Stock	Reads Selective Fields	Off Line	Punched Cards	
Standard Journal Tapes 1.31 to 3-1/4	32	15# to 20# Cal. .0025"-.0045"	Format Control by Computer, Limited Field Selectivity	On Line	Data to Computer	
2-1/4 x 3 to 5.91 to 9	85	20# to 100#	Format Control by 360 Computer Reads Selective Fields	On Line with IBM 360 Series	Data to Computer	Reads Mark Sense Documents Handprinted Digits & 3/16 Consecutive Numbers Serial Numbering of Doc.
3 x 6-1/2 to 9 x 14	81	16# to 100#	Format Control by Computer Reads Selective Fields	On Line with IBM 360 Series	Data to Computer	Reads Mark Sense Documents Handprinted Digits & 3/16 Consecutive Numbers Serial Numbering of Pages
2-3/4 x 5-7/8 to 3.67 x 8-3/4	80	Models 1 & 2 20# to 100# Model 3 20# to 125#	Reads Selective Fields	On Line to IBM 1400 Series & 360 Series Computers	Data to Computer	Reads Mark Sense Documents
3-1/2 x 2-1/4 to 8-3/4 x 4-1/4	80	Models 1 & 2 20# to 100# Model 3 20# to 125#	Reads Selective Fields	On Line to IBM 1400 Series & 360 Series Computers	Data to Computer	Reads Mark Sense Documents
5 x 3-1/2 to 8 x 3-1/2	72	20#, 24# or 100#	Reads Selective Fields	Off Line	Punched Cards Punched Paper Tape	Data Transmission
Standard Journal Tape 1.31 x 3-1/4	32	NCR Recommends Their 2AM3 Paper Rolls	Format Control Editing and Field Selection by Plugboard	On or Off Line with NCR, IBM 1400 Series and Univac 9000 Series	Data to Computer Tab Cards Punched Paper Tape Magnetic Tape	Header Line Entry
8-1/2 x 11	2840 Response Positions Available	60# Special Paper	Reads Selective Fields	Off Line	Punched Cards or Tape Magnetic Tape	
2-1/2 x 2-1/2 to 8-1/2 x 4-1/2	80 (Mach) 25 (Hand)	Specs not received from manufacturer.	Reads Intermixed or Selective Fields - Programmed by Plugboard	Off Line	Magnetic Tape, 7 or 9 Track, 550/ 800 bpi	
Optional Size Range Available (Depend on Model)	75	20# to 125#	Selective Fields; Extensive Formating and Editing Features	Off Line	Magnetic Tape Punched Cards or Paper Tape, or Data to Computer	Mark Reading Header Documents can be used for format specifications to program
2-1/4 x 4 to 2-1/4 x 8-1/2	80	20# to 125#	Limited Field Selectivity by External Computer	On Line	Data to Computer	
3-1/4 x 3-1/4 to 5 x 8-3/4	90	12# to 125#	Formating and Editing by Computer Reads Intermixed Fonts and Selective Fields	Off Line	Printer Punched Cards or Tape Magnetic Tape	Reads Mark Sense and Bar Codes, Accum. Totals
3-1/4 x 3-1/4 to 14 x 14	150	16# to 32#	Formating and Editing by Computer Reads Intermixed Fonts and Selective Fields	Off Line	Printer Punched Cards or Tape Magnetic Tape	Mark Reading and Bar Codes, Accum. Totals
Standard 80 Column Tab Card	40	100# Tab Card Stock	Reads Selective Fields - Programmed by Plugboard	Off Line	Punched Cards	
6-1/2 x 8 to 11 x 14	96	15# to 32#	Reads Selective Fields - Formating & Editing by Computer	On Line to Small General Purpose Computer	Data to Tape in General Purpose Computer	Reads Journal Tape as Optional Feature Handprint - 10 numeric and 10 symbols
Standard 51 or 80 Tab Cards, 3-1/4 to 8½ x Upward	80	20# to 125#	Reads Selective Fields	Off Line	Punched Paper Tape Data Transmission	Read Punched Hollerith Code
Standard 51 or 80 Column Tab Cards	80	100# Tab Card Stock	Reads Selective Fields	Off Line	Data Transmission	Read Punched Hollerith Code

ATTACHMENT 3.